Machine learning methods for pK_a prediction of small molecules: Advances and challenges

Jialu Wu, Yu Kang, Peichen Pan *, Tingjun Hou *

Innovation Institute for Artificial Intelligence in Medicine of Zhejiang University, College of Pharmaceutical Sciences and Cancer Center, Zhejiang University, Hangzhou 310058, Zhejiang, China

The acid–base dissociation constant (pK_a) is a fundamental property influencing many ADMET properties of small molecules. However, rapid and accurate pK_a prediction remains a great challenge. In this review, we outline the current advances in machine-learning-based QSAR models for pK_a prediction, including descriptor-based and graph-based approaches, and summarize their pros and cons. Moreover, we highlight the current challenges and future directions regarding experimental data, crucial factors influencing pK_a and *in silico* prediction tools. We hope that this review can provide a practical guidance for the follow-up studies.

Keywords: pK_a prediction; QSAR; Machine learning; Handcrafted features; Graph neural networks

Introduction

The acid–base dissociation constant (*pK$_a$*) is a key physicochemical parameter to describe the extent of proton dissociation reactions. K_a and its logarithmic form for a monoprotic acid (HA) are expressed in Equation (1):

$$pK_a(HA) = -\log_{10}K_a = -log_{10}\frac{[H^+][A^-]}{[HA]} \tag{1}$$

Conventionally, the pK_a of a basic compound is referred to that of its conjugate acid. For multiprotic compounds (compounds containing multiple ionizable centers), macro-pK_a and micro-pK_a need to be distinguished. In detail, micro-pK_a considers the loss or gain of a proton from a specific ionization site, whereas macro-pK_a reflects the dissociation ability of the whole molecule and is the net result of the equilibration of various microstates[1] (Equation (2)):

$$K_a^{macro} = \sum_{j=1}^{M^{deprot}} \frac{1}{\sum_{i=1}^{N^{prot}} \frac{1}{K_{ij}^{micro}}} \tag{2}$$

where M^{deprot} denotes M deprotonated microstates and N^{prot} denotes N protonated microstates.

pK_a is a fundamental physicochemical parameter widely applied in medicinal chemistry, organic synthesis, biochemistry, environmental science and materials science.[1] In drug discovery, pK_a determines the predominant protonation form of a drug-like molecule under specific tissues and organs with varied pH ranges, and thus has a high impact on its biological activity, ADMET profile[2,3] and other properties.

As an indispensable complement to experimental techniques, *in silico* pK_a prediction is more efficient and comparatively cheaper. Furthermore, theoretical calculations could provide fine-grained information inaccessible from experiments; for instance, the micro-pK_a of a single titratable site and even the nano-pK_a of a specific conformation.[4] Existing approaches for pK_a prediction can be divided into two categories: physics-based and empirical methods. The latter includes linear free energy relationship (LFER) and QSAR models. Physics-based pK_a prediction can be formulated as Equation (3), which relies on reaction free energy calculation and is commonly supplemented with linear empirical corrections (LECs) to absorb systematic errors[5]:

$$pK_a(HA) = A\frac{\Delta G_{aq}^{*}}{RTln(10)} + B \tag{3}$$

* *Corresponding authors.Pan, P. (panpeichen@zju.edu.cn), Hou, T. (tingjunhou@zju.edu.cn).*

1359-6446/© 2022 Elsevier Ltd. All rights reserved.
https://doi.org/10.1016/j.drudis.2022.103372

where ΔG^*_{aq} is the aqueous-phase reaction free energy at the standard state (T = 298.15 K, C = 1 mol/L), A and B are empirical constants derived from experimental data and R and T denote the gas constant and absolute temperature, respectively. The majority of the methods submitted to the SAMPL6 and SAMPL7 blind-challenge belongs to the physics-based category,[6] including *ab initio* quantum mechanics (QM),[7] density functional theory (DFT),[8–10] hybrid QM and molecular mechanics (MM),[11] embedded cluster reference interaction site model (EC-RISM),[12,13] integral equation formalism of the MiertusScrocco–Tomasi (IEFPCM/MST) model,[14] among others. However, the accuracy of solvation models remains a bottleneck[15] and the expensive computational cost is infeasible for high-throughput evaluation.

Assuming that the contributions of the substituents in a given class of acid or base are additive, LFER-based pK_a calculation can be parameterized by the Hammett–Taft (HT) equation[16]:

$$pK_a = pK_a^0 - \rho\sum_i^m \sigma_i \quad (4)$$

where pK_a^0 is the parent compound pK_a, ρ and σ are constants describing the substitution effect and m is the number of substituents. Empirical pK_a estimation using LFER has been extensively applied in commercial software, for instance, ACD/labs (https://www.acdlabs.com/products/percepta-platform/physchem-suite/pka/) and Epik (https://www.schrodinger.com/products/epik). Nevertheless, these models are confined to specific chemical series with available empirical constants for the parent structure and substituents.[17]

Over the past decade, QSAR modeling based on machine learning (ML) techniques has achieved remarkable success in pK_a prediction. Concretely, QSAR approaches can be classified into descriptor-based and graph-based models (Fig. 1). Descriptor-based models take the fixed-dimensional feature vectors extracted by human experts as input. By contrast, graph-based models operate directly on the molecular graphs annotated with basic atom and bond attributes. In this review, we outline the recent advances in ML-based pK_a prediction of small molecules, summarize the *in silico* prediction tools and discuss the current challenges and future directions.

Descriptor-based models

The prediction accuracy of descriptor-based models depends on the choice of input features. Currently, thousands of quantitative descriptors and qualitative fingerprints are available to represent small molecules.[18] For pK_a prediction, atomic descriptors and rooted fingerprints have been widely used to represent the local environment around the ionizable center. These local representations can be naturally applied to predict micro-pK_a.

Atomic descriptors

In this review, those real-valued parameters describing a certain atom or chemical bond are regarded as atomic descriptors. Most atomic descriptors derived from QM can be obtained by *ab initio*, DFT or semi-empirical[19] calculations. For a rapid estimation, they can also be obtained via empirical methods, such as Pauling's electronegativity and Gasteiger partial charge.[20] Existing work has validated the relationship between pK_a values and various atomic descriptors, including partial atomic charges,[21] Fukui frontier molecular orbital (FMO) descriptors (among which electrophilic superdelocalizability (SE) is the most contributive),[22] group philicity index,[23] quantum topological molecular similarity (QTMS) descriptors,[24] density functional reactivity theory (DFRT) descriptors [molecular electrostatic potential (MEP) and natural atomic orbital (NAO)],[25] among others.

The small number of atomic descriptors and their clear physicochemical meaning make atomic-descriptor-based models easily interpretable. The prior knowledge of the dissociation process can guide the choice of atomic descriptors and the model performance can lead to a better understanding of the chemical phenomena in turn. For instance, Popelier's group made an essential contribution to pK_a prediction based on the descriptor of *ab initio* bond lengths (AIBL). Their study demonstrated the strength of the AIBL-based models in handling tautomerizable molecules, which can somewhat overcome the lack of knowledge

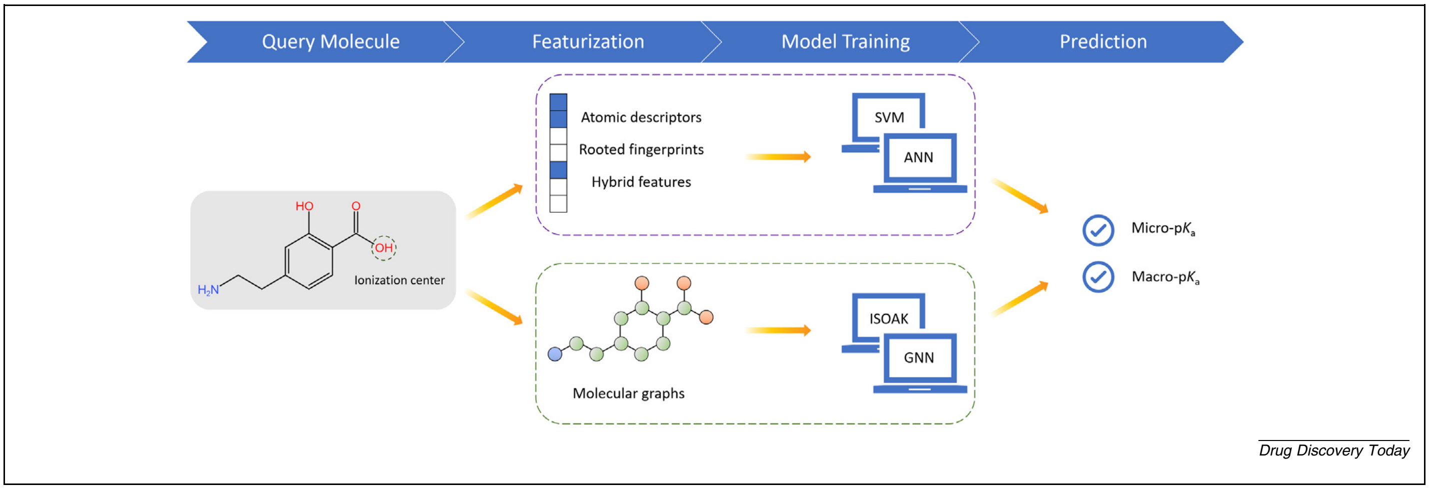


FIGURE 1

Overview of machine-learning-based QSAR modeling for pK_a prediction. The purple and green dashed boxes refer to descriptor-based and graph-based models, respectively. Abbreviations: SVM, support vector machine; ANN, artificial neural networks; ISOAK, iterative similarity optimal assignment kernel; GNN, graph neural networks.

on relative tautomeric stability. A single AIBL–pK_a relationship identified by statistical analysis can reduce computational cost and reveal the dominant tautomer.[26] They also highlighted the potential of AIBL for amending and augmenting experimental pK_a data.[27,28]

Typically, the atomic descriptors are calculated for the atoms of interest (AOI), including the ionization center undergoing protonation or deprotonation, the leaving proton[29] and other neighboring atoms (e.g., the atoms within the ionizable functional group).[23] The ML methods for modeling descriptor–pK_a relationships can range from simple linear regression to complex neural networks. Generally speaking, splitting molecules into subsets allows a simpler model with fewer descriptors.[22] However, the class-specific models can suffer from limited application domain and high risk of overfitting.

To address the above issues, Skolidis *et al.*[30] successfully implemented multitask learning to improve model performance on some subclasses with insufficient data by utilizing the data from related tasks. Hunt *et al.*[19] took a step forward by building a versatile model to predict pK_a for diverse mono- and multi-protic compounds using radial basis function (RBF). The main idea of their model is to use semi-empirical QM descriptors to capture atomic and bond properties in the forms of conjugate acid and conjugate base, which considers the whole molecule environment with acceptable computational cost. Similarly, Bannan *et al.*[31] built a general Gaussian process (GP) model based on ten physical features describing a specific form or the difference between two conjugated forms. Particularly, the GP model can provide an uncertainty estimation reflecting its confidence of the prediction. In 2021, Raddi *et al.*[32] constructed a deep GP model by utilizing more features, which achieved significant improvements over the standard GP model.

Rooted fingerprints

Rooted fingerprints are binary or count-based feature vectors to describe the local structural environment around the specified root atom. Here, the root atom refers to the ionization atom. A pioneering study by Xing *et al.*[33] introduced a novel molecular-tree-structured fingerprint to describe the composition of atom and/or group types at each level originated from the ionization center. In addition, creating tailored fingerprints for each functional group could considerably improve model performance.

In MoKa, Milletti *et al.*[34] proposed a fingerprint derived from GRID molecular interaction fields (MIFs). The model based on a fragment database precomputed by MIFs can rapidly describe atoms using energy minima, then convert them into one-hot vectors (binned by energy values) and sum up at each topological distance. A strong advantage of this method is that the MIF-derived parameters can encode 3D information to some extent, for instance H-bonding and steric effects.

Lee *et al.*[35] created a decision tree model (SMARTS pK_a) based on a novel set of SMARTS strings, which discriminates different ionization centers and substituents. In this way, they avoided overfitting caused by dividing training data into class-specific subsets, and the prediction results would be a pK_a range determined by the training samples sharing the same leaf node.

In 2019, Lu *et al.*[36] employed rooted topological torsion fingerprints (RTorsion), which explicitly encode the path and atom connectivity information, for pK_a prediction of aliphatic amines. Specifically, a 1-bond path can be described as [(N, 0, 0), (C, 1, 0)], which includes the information of atom types, number of non-hydrogen atom neighbors not in this path and the number of π election pairs. The results proved that RTorsion coupled with ML methods, especially support vector machine (SVM), can provide excellent prediction (RMSE = 0.45, MAE = 0.33 and R^2 = 0.84 on the external test set with 726 pK_a values).

In 2021, Plante *et al.*[28] designed distance spectrum fingerprints where each feature reflects the impact of a specific atom type on the pK_a values, similar to the molecular-tree-structured fingerprints. Subsequently, atom-type coefficients ($\alpha_{Atom-type}$ in Equation (5)) were generated via partial least squares (PLS). The special design is that they explicitly consider the topological distance decay effect. Moreover, the six-digit number (in the format of ABBCDD) representing atom types is highly extendable.

$$pK_a = \sum_{All\ atoms} \frac{\alpha_{Atom-type}}{Topological\ Distance^2_{atom}} \tag{5}$$

Hybrid features

In recent years, researchers have extensively experimented with various combinations of molecular features and ML algorithms and made significant contributions in terms of open-source data, codes and prediction tools. Descriptors and fingerprints can supplement each other because descriptors emphasize physicochemical properties whereas fingerprints focus on structural information.[20,36] Mansouri *et al.*[37] utilized continuous molecular descriptors, binary fingerprints and fragment counts generated by PaDEL to construct pK_a prediction models and proved that the hybrid features consistently outperformed a single feature set. In particular, they separated molecules into the acidic and basic subsets and built the prediction models separately. The SVM model combined with a k-nearest neighbor (kNN) classifier performed best and was implemented in OPERA.[38] Baltruschat *et al.*[39] comprehensively evaluated six combinations of molecular features and found that the random forest (RF) model based on the RDKit descriptors and extended connectivity fingerprints (ECFP) achieved the best predictions. In the external tests, the RF model beat OPERA, possibly owing to the expanded training data and the choice of building a single model to handle compounds with their protonated states at pH = 7.4. In 2021, Yang *et al.*[1] pioneeringly developed a holistic model capable of simultaneously predicting aqueous and nonaqueous pK_a by fully utilizing the intrinsic relationship of pK_a values between solvents. They introduced structural and physical-organic-parameter-based descriptors (SPOC), which contain the RDKit descriptors and MACCS fingerprints, combined with a novel ionic status labeling (ISL) to distinguish three pK_a subtypes (i.e., pK_a values of neutral, protonated and negatively charged molecules). The model trained with extreme gradient boosting (XGBoost) or artificial neural networks (ANN) algorithms achieved a low MAE of 0.87 pK_a.

By contrast, local representations can better capture electronic effects but might ignore remote effects and the interactions between multiple dissociation sites. Furthermore, taking the global representations into account can heuristically help ML methods to better determine the relative importance of each atomic descriptor.[40] The representative work by combining local and global representations includes: (i) rooted ECFP and standard ECFP[20]; (ii) atomic descriptors and size-related descriptors[41]; and (iii) atomic descriptors and standard ECFP.[32] It is important to stress that global representations can also introduce confusing information about distant and irrelevant functional groups.[20].

Graph-based models

Small molecules can be naturally described as graphs in which nodes and edges denote atoms and chemical bonds, respectively. Graph-based methods can extract information directly from the annotated molecular graphs, demonstrating a great advantage in molecular property prediction.

Graph kernels

Conventionally, graph-structured data can be represented using graph kernels, which compute an inner product on graphs to measure their similarity. Graph kernels allow the kernelized ML methods to work directly on graphs without intermediate conversion from graphs to feature vectors, thus avoiding the loss of structural information.[42] In 2010, Rupp *et al.*[43] employed iterative similarity optimal assignment kernel (ISOAK) and kernel ridge regression (KRR) to estimate pK_a values. The results demonstrated that the graph-kernel-based methods could yield comparable performance compared with the semi-empirical models based on frontier electron theory, and time-consuming structure optimization is unnecessary. Significantly, the graph kernel approach performed better on larger series of compounds with high structural diversity, showing its potential for developing a generic model.

Graph neural networks

Although graph kernels can directly operate on graphs, they are still manually engineered and cannot learn the optimal representations tailored to the downstream tasks. In this context, scientists extended the convolutional neutral networks (CNN) to graph-structured data and proposed graph neural networks (GNN). In 2015, Duvenaud *et al.*[44] first introduced GNN to learn the differentiable neural graph fingerprints (Neural FP) with end-to-end supervision. Neural FP is a type of real-valued vector where each feature can be activated by similar but distinct fragments, making the representations more meaningful and low-dimensional. Extensive experiments demonstrated that Neural FP exhibited stronger predictive power and interpretability than circular fingerprints. In 2016, Kipf *et al.*[45] formally introduced the concept of graph convolutional networks (GCN). Subsequently, a series of GNN variants were proposed, including message passing neural networks (MPNN),[46] graph attention networks (GAT),[47] directed MPNN (D-MPNN),[48] Attentive FP,[49] among others.

In the past few years, GNN has attracted growing interest and achieved considerable success in drug discovery, including molecular property prediction,[50] *de novo* drug design,[51] drug–target interactions,[52] and so on. Without exception, the effectiveness of GNN in the pK_a prediction task has also been evaluated. Given that pK_a is an atom-centered property, it can be treated as either a node-level or graph-level task. Dealing with node-level tasks, GNN utilizes stacked graph convolution layers to aggregate the information from neighborhoods and update the hidden states of the target node (message passing phase). As for graph-level tasks, a global pooling operation is implemented to summarize node-level representations into a graph-level representation (readout phase).

In 2019, Roszak *et al.*[20] took the lead in employing GCN to estimate the acidity of C—H groups in nonaqueous solvents. To find the most acidic proton in the molecule, they directly used the node embeddings to predict the pK_a values for all atoms (Fig. 2a). The work evidenced that GCN can provide rapid (within milliseconds) and accurate predictions for atom-specific characteristics. In realistic chemical examples, the proposed GCN model correctly predicted the reacting site in > 90 % of cases, showing its potential application in synthetic planning.

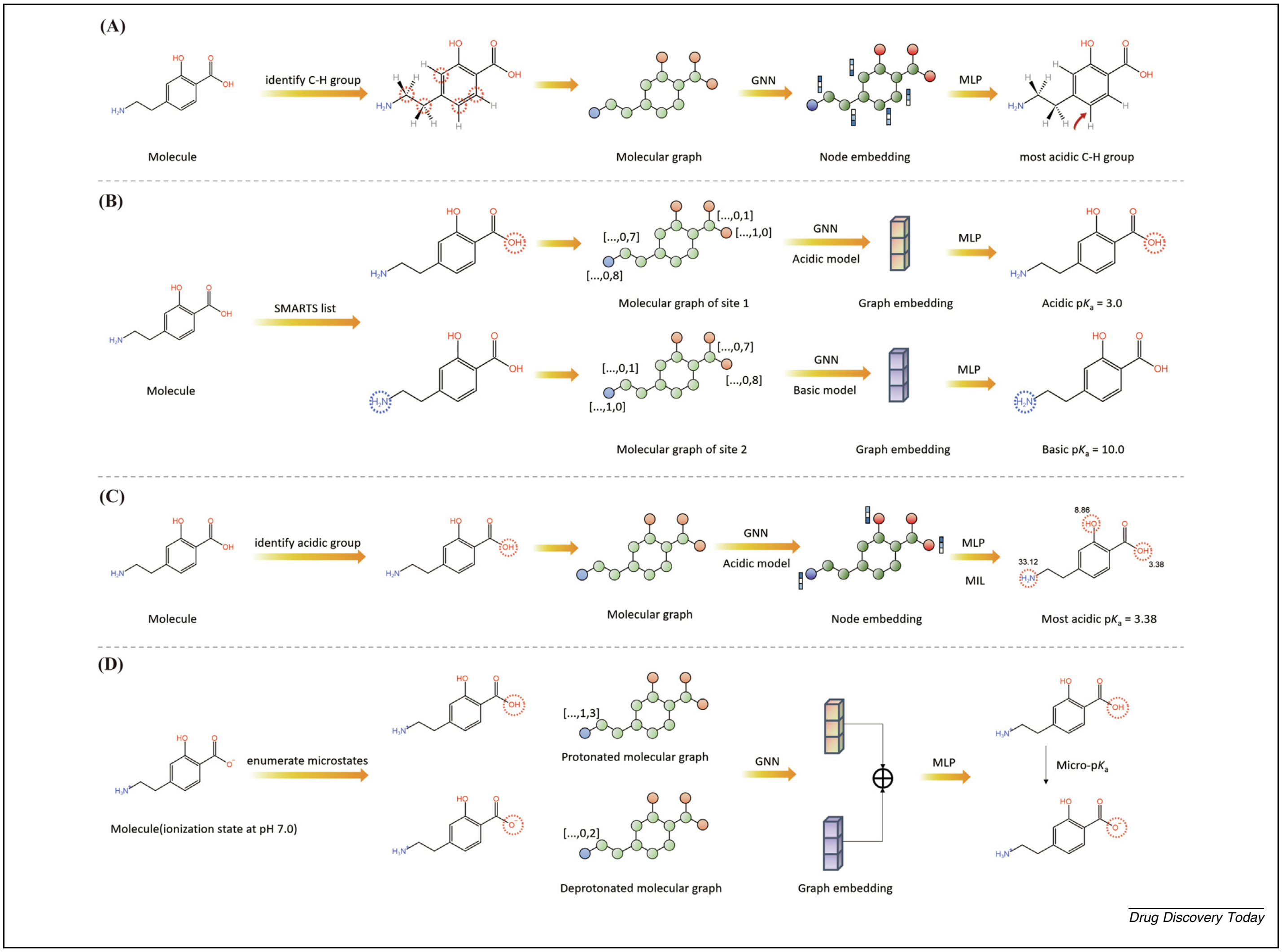
The follow-up studies incorporated prior domain knowledge into the GNN architectures for pK_a prediction, which introduced inductive bias and improved the model's ability to capture task-related information. In 2021, Pan *et al.*[53] developed a GCN-based pK_a predictor named MolGpK_a. They added two extra dimensions in the initial atom features to encode the ionization centers, one is a binary flag and the other is the shortest topological distance to the target center. Given a multiprotic molecule, they can identify the ionization sites via substructure matching and get the graph representations for each specified site one by one (Fig. 2b). They extended the SMARTS list provided by Ropp *et al.*[54] to cover all oxygen, nitrogen and sulfur centers in the training set, finally containing 144 ionizable groups. The detailed analysis showed that MolGpK_a achieved comparable performance to commercial software Marvin (Table 1) and Epik and the learned effects of substituents agreed well with expert knowledge.

Later, Xiong *et al.*[55] combined multi-instance learning (MIL) and Attentive FP to establish a novel model capable of predicting micro-pK_a and macro-pK_a called Graph-pK_a. Similar to the work by Roszak *et al.*,[20] Graph-pK_a predicts the micro-pK_a directly from the learned node features. Unlike the previous work that obtains micro-pK_a from QM calculation[20] or assigns the macro-pK_a to the dominant site as an approximation,[1,19,53] which will inevitably introduce data noise, Graph-pK_a provides a new paradigm for dealing with micro-pK_a. In detail, Graph-pK_a follows MIL to calculate macro-pK_a (label of bags) from predicted micro-pK_a (label of instance) utilizing the approximate mathematical relationships (Equations (6) and (7)) between them, thus training against the experimental macro-pK_a labels (Fig. 2c). Graph-pK_a achieved state-of-the-art performance on the SAMPL6 dataset and exhibited indistinguishable intelligence from a human expert in locating the most acidic and basic sites of molecules (evaluated by consistency rate and difference values).

$$\mathrm{p}K_{\mathrm{a(acidic)}} = -\log\left(\sum_{i=1}^{N} 10^{-\mathrm{p}K_{\mathrm{a(acidic)}}{}^{i}}\right) \qquad (6)$$

$$\mathrm{p}K_{\mathrm{a(basic)}} = \log\left(\sum_{i=1}^{N} 10^{\mathrm{p}K_{\mathrm{a(basic)}}{}^{i}}\right) \qquad (7)$$

A limitation of MolGpK_a and Graph-pK_a is that they can only deal with neutral molecules and predict the most acidic and basic pK_a values. More recently, Mayr *et al.*[56] developed a workflow named p$K_a$solver to realize sequential pK_a prediction. To this end, they identified the ionizable sites of a given compound using Dimorphite-DL,[54] and then generated protonated-



Drug Discovery Today

FIGURE 2

Illustration of GNN-based pK_a prediction models. **(a)** Roszak's model.[20] **(b)** MolGpK_a[53]: the atom features refer to the ionization center flag and the shortest distance to the target center. **(c)** Graph-pK_a.[55] **(d)** p$K_a$solver[56]: the atom features refer to the formal charge and the total number of hydrogens. The red and blue circles refer to acidic and basic ionization centers, respectively. Abbreviations: GNN, graph neural networks; MLP, multilayer perceptron.

TABLE 1

QSAR-based *in silico* tools for p*K*a prediction.

Name	Molecular representation	Methods	URL	Availability
Marvin[57]	Atomic descriptors	MLR	https://chemaxon.com/products/marvin	Commercial
ADMET redictor (S + pK_a)[58]	Atomic descriptors	Ensemble of ANN	https://www.simulations-plus.com/software/admetpredictor/	Commercial
MoKa[22,59]	Rooted fingerprints	PLS	https://www.moldiscovery.com/software/moka/	Commercial
OPERA[25]	Hybrid features	kNN + SVM	https://github.com/NIEHS/OPERA	Free
Yang's model[1]	Hybrid features	ANN/XGBoost	https://pKa.luoszgroup.com	Free
Baltruschat's model[27]	Hybrid features	RF	https://github.com/czodrowskilab/Machine-learning-meets-pKa	Free
Roszak's model[10]	Molecular graphs	GNN	https://pKa.allchemy.net	Free
MolGpK_a[42]	Molecular graphs	GNN	https://xundrug.cn/molgpKa	Free
Graph-pK_a[44]	Molecular graphs	GNN	https://pKa.simm.ac.cn	Free
p$K_a$solver[45]	Molecular graphs	GNN	https://github.com/167mayrf/pKasolver	Free

Abbreviations: MLR, multiple linear regression; ANN, artificial neural networks; PLS, partial least squares; kNN, k-nearest neighbor; SVM, support vector machine; XGBoost, extreme gradient boosting; RF, random forest; GNN, graph neural networks.

deprotonated pairs where graph embeddings are concatenated for micro-pK_a prediction (Fig. 2d). To reproduce the predominant dissociation process, the microstates enumeration is achieved starting from the structure at pH = 7.0 and in an iterative manner, implying that the most acidic or basic site will be ionized for the next round calculation.

Challenges and future directions

The accurate prediction of pK_a remains a challenging problem owing to the data scarcity and the intrinsic complexity of the property. Currently, the quantity and quality of open-source data are unsatisfactory. The freely available database DataWarrior[57] records the pK_a values for 7912 chemicals but the amount of the valid data after curation is reduced to 6188. If acidic pK_a and basic pK_a are treated separately, the available data for each model is only ~ 3000.[37] To expand the modeling data, researchers adopted time-consuming QM calculation[20] or tedious manual collection from previous work and literature.[55] Xiong *et al.* compiled a large S-pK_a dataset containing 16 595 compounds with 17 489 pK_a values but it is not publicly available.[55] The authors also pointed out that the difficulty of collecting and labeling multi-step pK_a data hinders the development of prediction models. The *i*BOND database (https://ibond.nankai.edu.cn/) established by Tsinghua and Nankai university is the largest academic database, providing > 30 000 experimental pK_a data in 39 solvents with major ionization sites assigned by experts.[1] The database is easy to search but cannot be downloaded for model training. In this case, leveraging computational pK_a values[53] or enhancing the collaboration between academia and industry[58] are feasible avenues. Furthermore, benchmark datasets comprising novel and structurally diverse compounds related to real-life drug discovery are required for comparative assessment of prediction tools. The SAMPL blind challenges[2,59] represent a solid start but the data are still limited.

Another challenge is that complicated factors affect the ionization of a particular group, including conformational flexibility,[60] structural symmetry,[60] unusual heterocycles,[60] multiple ionization centers,[61] charge transfer in conjugated systems,[61] tautomerism[62] and intra- or inter-molecular interactions (hydrogen and halogen bonding[15]). One potential solution is to couple domain knowledge and state-of-the-art deep learning (DL) algorithms (such as 3D-informed GNN[63,64]) to better capture chemical patterns from training data.

Table 1 summarizes the QSAR-based *in silico* tools for pK_a prediction. We can conclude the current trends as follows: (i) a growing number of open-source prediction tools are emerging; (ii) data-driven features are gradually replacing handcrafted features; (iii) more attention is shifting from establishing class-specific models to building generic models. Compared with commercial software, open-source tools still have major deficiencies. First, they are unable to comprehensively solve three pK_a-related tasks, namely micro-pK_a prediction, macro-pK_a prediction and the proportions of each microstate under different pH conditions.[6] Second, the automation degree remains to be improved, mainly reflected in tautomer enumeration, multi-step pK_a prediction and batch evaluation. Finally, limited attention is paid to model interpretation, although it is a significant step for rationalizing predictions to convince chemical scientists. Besides, it might be instrumental in uncovering previous unknown chemical knowledge and guiding lead optimization. We hope the follow-up studies can be targeted to overcome the above issues.

Concluding remarks

The *in silico* prediction of pK_a has a profound impact on chemical science, especially in drug discovery. In the past decades, researchers have made remarkable progress in the field of pK_a prediction and developed various accessible tools. However, there are still multiple issues that remain to be solved, mainly attributed to insufficient data and the intricate effects derived from structural factors.

As we know, the predictive power of QSAR-based models highly depends on the quality of the input features. Fortunately, the advent of GNN provides an intelligent solution to learning the expressive features directly from molecular graphs. The superiority of GNN in pK_a prediction tasks, whether micro- or macro-pK_a, has been well confirmed in recent studies. In our opinion, exploiting the synergy of expert knowledge and GNN architecture has the potential to capture more structure–pK_a relationships from less data, thus overcoming the abovementioned bottlenecks and constructing more-reliable models.

It is indispensable that we should comprehensively evaluate the predictive ability and application domain of the current approaches, which is beneficial to model selection and improvement. We believe that the accumulation of high-quality data and the emergence of powerful algorithms will enable the development of accurate, efficient, versatile and interpretable pK_a prediction models. Regarding future applications, pK_a prediction tools can be utilized as a plugin in the workflow of artificial intelligence (AI)-driven drug discovery; for example, property optimization in molecule generation and property filter in virtual screening. Thus, the models proposed by the scientific community can play a part in manufacturing practice and reduce the risk of failure veritably.

Data availability

No data was used for the research described in the article.

Acknowledgments

This work was financially supported by National Key Research and Development Program of China (2021YFF1201400), Natural Science Foundation of Zhejiang Province (LZ19H300001), Key R&D Program of Zhejiang Province (2020C03010) and Fundamental Research Funds for the Central Universities (2020QNA7003).

References

1 Q. Yang, Y. Li, J.D. Yang, et al., Holistic Prediction of the pKa in Diverse Solvents Based on a Machine-Learning Approach, Angew Chem Int Ed 59 (2020) 19282–19291.

2 T.D. Bergazin, N. Tielker, Y. Zhang, et al., Evaluation of log P, pKa, and log D predictions from the SAMPL7 blind challenge, J Comput Aided Mol Des 35 (2021) 771–802.

3 A. Avdeef, Prediction of aqueous intrinsic solubility of druglike molecules using random forest regression trained with Wiki-pS0 database, ADMET DMPK 8 (2020) 50.

4 A.D. Bochevarov, M.A. Watson, J.R. Greenwood, D.M. Philipp, Multiconformation, Density Functional Theory-Based pKa Prediction in Application to Large, Flexible Organic Molecules with Diverse Functional Groups, J Chem Theory Comput 12 (2016) 6001–6019.

5 A. Klamt, F. Eckert, M. Diedenhofen, M.E. Beck, First Principles Calculations of Aqueous pKa Values for Organic and Inorganic Acids Using COSMO−RS Reveal an Inconsistency in the Slope of the pka Scale, J Phys Chem A 107 (2003) 9380–9386.

6 M. Iş k, A.S. Rustenburg, A. Rizzi, M.R. Gunner, D.L. Mobley, J.D. Chodera, Overview of the SAMPL6 pKa challenge: evaluating small molecule microscopic and macroscopic pKa predictions, J Comput Aided Mol Des 35 (2021) 131–166.

7 E. Selwa, I.M. Kenney, O. Beckstein, B.I. Iorga, SAMPL6: calculation of macroscopic pKa values from ab initio quantum mechanical free energies, J Comput Aided Mol Des 32 (2018) 1203–1216.

8 Q. Zeng, M.R. Jones, B.R. Brooks, Absolute and relative pKa predictions via a DFT approach applied to the SAMPL6 blind challenge, J Comput Aided Mol Des 32 (2018) 1179–1189.

9 P. Pracht, R. Wilcken, A. Udvarhelyi, S. Rodde, S. Grimme, High accuracy quantum-chemistry-based calculation and blind prediction of macroscopic pKa values in the context of the SAMPL6 challenge, J Comput Aided Mol Des 32 (2018) 1139–1149.

10 B.K. F nd k, Z.P. Haslak, E. Arslan, V. Aviyente, SAMPL7 blind challenge: quantum–mechanical prediction of partition coefficients and acid dissociation constants for small drug-like molecules, J Comput Aided Mol Des 35 (2021) 841–851.

11 S. Prasad, J. Huang, Q. Zeng, B.R. Brooks, An explicit-solvent hybrid QM and MM approach for predicting pKa of small molecules in SAMPL6 challenge, J Comput Aided Mol Des 32 (2018) 1191–1201.

12 N. Tielker, L. Eberlein, S. Güssregen, S.M. Kast, The SAMPL6 challenge on predicting aqueous pKa values from EC-RISM theory, J Comput Aided Mol Des 32 (2018) 1151–1163.

13 N. Tielker, S. Güssregen, S.M. Kast, SAMPL7 physical property prediction from EC-RISM theory, J Comput Aided Mol Des 35 (2021) 933–941.

14 A. Viayna, S. Pinheiro, C. Curutchet, F.J. Luque, W.J. Zamora, Prediction of n-octanol/water partition coefficients and acidity constants (pKa) in the SAMPL7 blind challenge with the IEFPCM-MST model, J Comput Aided Mol Des 35 (2021) 803–811.

15 P.G. Seybold, G.C. Shields, Computational estimation of pKa values, WIREs Comput Mol Sci 5 (2015) 290–297.

16 M. Rupp, R. Korner, V. Tetko I., Predicting the pKa of Small Molecules, Comb Chem High Throughput Screen 14 (2011) 307–327.

17 Fraczkiewicz R. In Silico Prediction of Ionization. In: *Comprehensive Medicinal Chemistry II*. Vol 5; 2006:603–26.

18 K.V. Chuang, L.M. Gunsalus, M.J. Keiser, Learning Molecular Representations for Medicinal Chemistry: Miniperspective, J Med Chem 63 (2020) 8705–8722.

19 P. Hunt, L. Hosseini-Gerami, T. Chrien, J. Plante, D.J. Ponting, M. Segall, Predicting pKa Using a Combination of Semi-Empirical Quantum Mechanics and Radial Basis Function Methods, J Chem Inf Model 60 (2020) 2989–2997.

20 R. Roszak, W. Beker, K. Molga, B.A. Grzybowski, Rapid and Accurate Prediction of pKa Values of C-H Acids Using Graph Convolutional Neural Networks, J Am Chem Soc 141 (2019) 17142–17149.

21 Z.P. Haslak, S. Zareb, I. Dogan, V. Aviyente, G. Monard, Using Atomic Charges to Describe the p K_a of Carboxylic Acids, J Chem Inf Model 61 (2021) 2733–2743.

22 B.G. Tehan, E.J. Lloyd, M.G. Wong, et al., Estimation of pKa Using Semiempirical Molecular Orbital Methods. Part 1: Application to Phenols and Carboxylic Acids, Quant Struct-Act Relatsh 21 (2002) 457–472.

23 R. Parthasarathi, J. Padmanabhan, M. Elango, K. Chitra, V. Subramanian, P.K. Chattaraj, pKa Prediction Using Group Philicity, J Phys Chem A 110 (2006) 6540–6544.

24 A.P. Harding, D.C. Wedge, P.L.A. Popelier, pKa Prediction from “Quantum Chemical Topology” Descriptors, J Chem Inf Model 49 (2009) 1914–1924.

25 Y. Huang, L. Liu, W. Liu, S. Liu, S. Liu, Modeling Molecular Acidity with Electronic Properties and Hammett Constants for Substituted Benzoic Acids, J Phys Chem A 115 (2011) 14697–14707.

26 B.A. Caine, M. Bronzato, T. Fraser, N. Kidley, C. Dardonville, P.L.A. Popelier, Aqueous pKa prediction for tautomerizable compounds using equilibrium bond lengths, Commun Chem 3 (2020) 21.

27 B.A. Caine, M. Bronzato, P.L.A. Popelier, Experiment stands corrected: accurate prediction of the aqueous p K_a values of sulfonamide drugs using equilibrium bond lengths, Chem Sci 10 (2019) 6368–6381.

28 J. Plante, B.A. Caine, P.L.A. Popelier, Enhancing Carbon Acid pKa Prediction by Augmentation of Sparse Experimental Datasets with Accurate AIBL (QM) Derived Values, Molecules 26 (2021) 1048.

29 S. Liu, C.K. Schauer, L.G. Pedersen, Molecular acidity: A quantitative conceptual density functional theory description, J Chem Phys 131 (2009) 164107.

30 G. Skolidis, K. Hansen, G. Sanguinetti, M. Rupp, Multi-task learning for pKa prediction, J Comput Aided Mol Des 26 (2012) 883–895.

31 C.C. Bannan, D.L. Mobley, A.G. Skillman, SAMPL6 challenge results from pKa predictions based on a general Gaussian process model, J Comput Aided Mol Des 32 (2018) 1165–1177.

32 R.M. Raddi, V.A. Voelz, Stacking Gaussian processes to improve pKa predictions in the SAMPL7 challenge, J Comput Aided Mol Des 35 (2021) 953–961.

33 L. Xing, R.C. Glen, R.D. Clark, Predicting pKa by Molecular Tree Structured Fingerprints and PLS, J Chem Inf Comput Sci 43 (2003) 870–879.

34 F. Milletti, L. Storchi, G. Sforna, G. Cruciani, New and Original pKa Prediction Method Using Grid Molecular Interaction Fields, J Chem Inf Model 47 (2007) 2172–2181.

35 A.C. Lee, J. Yu, yu, Crippen GM., pKa Prediction of Monoprotic Small Molecules the SMARTS Way, J Chem Inf Model 48 (2008) 2042–2053.

36 Y. Lu, S. Anand, W. Shirley, et al., Prediction of pKa Using Machine Learning Methods with Rooted Topological Torsion Fingerprints: Application to Aliphatic Amines, J Chem Inf Model 59 (2019) 4706–4719.

37 K. Mansouri, N.F. Cariello, A. Korotcov, et al., Open-source QSAR models for pKa prediction using multiple machine learning approaches, J Cheminformatics 11 (2019) 60.

38 K. Mansouri, C.M. Grulke, R.S. Judson, A.J. Williams, OPERA models for predicting physicochemical properties and environmental fate endpoints, J Cheminformatics 10 (2018) 10.

39 M. Baltruschat, P. Czodrowski, Machine learning meets pKa, *F1000Research* 9 (2020). Chem Inf Sci-113.

40 J. Zaretzki, M. Matlock, S.J. Swamidass, XenoSite: Accurately Predicting CYP-Mediated Sites of Metabolism with Neural Networks, J Chem Inf Model 53 (2013) 3373–3383.

41 R. Lawler, Y.H. Liu, N. Majaya, et al., DFT-Machine Learning Approach for Accurate Prediction of p K_a, J Phys Chem A 125 (2021) 8712–8722.

42 M. Rupp, E. Proschak, G. Schneider, Kernel Approach to Molecular Similarity Based on Iterative Graph Similarity, J Chem Inf Model 47 (2007) 2280–2286.

43 M. Rupp, R. Körner, I.V. Tetko, Estimation of Acid Dissociation Constants Using Graph Kernels, Mol Inform 29 (2010) 731–740.

44 Duvenaud D, Maclaurin D, Aguilera-Iparraguirre J, et al. Convolutional Networks on Graphs for Learning Molecular Fingerprints. *arXiv:150909292*. Published online November 3, 2015.

45 Kipf TN, Welling M. Semi-Supervised Classification with Graph Convolutional Networks. *ArXiv160902907 Cs Stat*. Published online February 22, 2017.

46 Gilmer J, Schoenholz SS, Riley PF, Vinyals O, Dahl GE. Neural Message Passing for Quantum Chemistry. *ArXiv170401212 Cs*. Published online June 12, 2017.

47 Veličković P, Cucurull G, Casanova A, Romero A, Liò P, Bengio Y. Graph Attention Networks. *ArXiv171010903 Cs Stat*. Published online February 4, 2018.

48 K. Yang, K. Swanson, W. Jin, et al., Analyzing Learned Molecular Representations for Property Prediction, J Chem Inf Model 59 (2019) 3370–3388.

49 Z. Xiong, D. Wang, X. Liu, et al., Pushing the Boundaries of Molecular Representation for Drug Discovery with the Graph Attention Mechanism, J Med Chem 63 (2020) 8749–8760.

50] O. Wieder, S. Kohlbacher, M. Kuenemann, et al., A compact review of molecular property prediction with graph neural networks, Drug Discov Today Technol 37 (2020) 1–12.

51 J. Xiong, Z. Xiong, K. Chen, H. Jiang, M. Zheng, Graph neural networks for automated de novo drug design, Drug Discov Today 26 (2021) 1382–1393.

52 Z. Zhang, L. Chen, F. Zhong, et al., Graph neural network approaches for drug-target interactions, Curr Opin Struct Biol 73 (2022) 102327.

53 X. Pan, H. Wang, C. Li, J.Z.H. Zhang, C. Ji, MolGpka: A Web Server for Small Molecule pKa Prediction Using a Graph-Convolutional Neural Network, J Chem Inf Model 61 (2021) 3159–3165.

54 P.J. Ropp, J.C. Kaminsky, S. Yablonski, J.D. Durrant, Dimorphite-DL: an open-source program for enumerating the ionization states of drug-like small molecules, J Cheminformatics 11 (2019) 14.

55 J. Xiong, Z. Li, G. Wang, et al., Multi-instance learning of graph neural networks for aqueous p K a prediction. Lu Z, ed, Bioinformatics 38 (2022) 792–798.

56 F. Mayr, M. Wieder, O. Wieder, T. Langer, *Improving Small Molecule PK*$_a$ *Prediction Using Transfer Learning with Graph Neural Networks*, Biophysics (2022).

57 T. Sander, J. Freyss, M. von Korff, C. Rufener, DataWarrior: An Open-Source Program For Chemistry Aware Data Visualization And Analysis, J Chem Inf Model 55 (2015) 460–473.

58 R. Fraczkiewicz, M. Lobell, A.H. Göller, et al., Best of Both Worlds: Combining Pharma Data and State of the Art Modeling Technology To Improve in Silico pKa Prediction, J Chem Inf Model 55 (2015) 389–397.

59 M. Iş k, D. Levorse, A.S. Rustenburg, et al., pKa measurements for the SAMPL6 prediction challenge for a set of kinase inhibitor-like fragments, J Comput Aided Mol Des 32 (2018) 1117–1138.

60 D.M. Philipp, M.A. Watson, H.S. Yu, T.B. Steinbrecher, A.D. Bochevarov, Quantum chemical prediction for complex organic molecules, Int J Quantum Chem 118 (2018) e25561.

61 S. Jelfs, P. Ertl, P. Selzer, Estimation of pKa for Druglike Compounds Using Semiempirical and Information-Based Descriptors, J Chem Inf Model 47 (2007) 450–459.

62 G. Cruciani, F. Milletti, L. Storchi, G. Sforna, L. Goracci, In silico pKa Prediction and ADME Profiling, Chem Biodivers 6 (2009) 1812–1821.

63 Stärk H, Beaini D, Corso G, et al. 3D Infomax improves GNNs for Molecular Property Prediction. *arXiv:211004126*. Published online October 8, 2021.

64 Fang X, Liu L, Lei J, et al. ChemRL-GEM: Geometry Enhanced Molecular Representation Learning for Property Prediction. *arXiv:210606130*. Published online July 29, 2021.